IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Survey of DNN Methods For Blind Image Quality Assessment

XIAOHAN YANG[1], (Student Member, IEEE), FAN LI[1], (MEMBER, IEEE), AND HANTAO LIU[2], (Member, IEEE)

[1]the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. (e-mail: yangxiaohan@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn)

[2]the School of Computer Science and Informatics, Cardiff University, Cardiff, CF243AA, U.K. (e-mail: LiuH35@cardiff.ac.uk)

Corresponding author: Fan Li (e-mail: lifan@mail.xjtu.edu.cn).

This research work was supported in part by National Science Foundation of China (61671365), and Joint Foundation of Ministry of Education of China (6141A02022344).

ABSTRACT Blind image quality assessment (BIQA) methods aim to predict quality of images as perceived by humans without access to a reference image. Recently, deep learning methods have gained substantial attention in the research community and have proven useful for BIQA. Although previous study of deep neural networks (DNN) methods is presented, some novelty DNN methods, which are recently proposed, are not summarized for BIQA. In this paper, we provide a survey covering various DNN methods for BIQA. First, we systematically analyze the existing DNN-based quality assessment methods according to the role of DNN. Then, we compare the prediction performance of various DNN methods on the synthetic databases (LIVE, TID2013, CSIQ, LIVE multiply distorted) and authentic databases (LIVE challenge), providing important information that can help understand the underlying properties between different DNN methods for BIQA. Finally, we describe some emerging challenges in designing and training DNN-based BIQA, along with few directions that are worth further investigations in the future.

INDEX TERMS deep learning, blind image quality assessment (BIQA), deep neural networks (DNN) model, deep features, quality prediction.

I. INTRODUCTION

WITH the development of social media and the increasing demand for imaging services, an enormous amount of visual data is making its way to consumers. Digital images are likely to be inevitably degraded in the processes from content generation to consumption. The acquisition, processing, compression, transmission, or storage of images is subject to various distortions, resulting degradation in visual quality. Therefore, methods for image quality assessment (IQA) have been extensively studied for the purpose of maintain, control and enhance the perceived image quality.

In principal, subjective assessment is the most reliable way to evaluate the visual quality of images [1], [2]. But this method is time-consuming, expensive, and difficult to implement in real-world systems. Therefore, objective assessment of image quality has gained growing attention in recent years. To what extent a reference image is used for quality assessment, existing objective IQA methods can be classified into three categories: full-reference (FR), reduced-reference (RR) and no-reference/blind (NR/B) methods. The FR IQA methods make full use of the undistorted reference images to compare with distorted images and measure the difference between them [3]–[5], while the RR IQA methods use partial information in reference images [6]–[8]. However, in many practical applications, it is difficult to obtain a reference image of the distorted image to be assessed, making powerful FR and RR IQA methods inapplicable. On the contrary, the BIQA methods have no access to the reference images to evaluate image quality [9], [10]. Thus, it has become increasingly important to develop effective BIQA methods which can predict image quality without any additional information.

Most exiting BIQA methods follow the flowchart shown in Fig. 1. Some BIQA methods is developed based on classical regression methods [11].Researchers attempt to design some hand-crafted features that could discriminate distorted images, and then train a regression model to predict image quality. Early BIQA methods are based on a distortion specific approach [78], [79], which commonly uses the prior knowledge of the distortion types for quality prediction. In this

approach, the distortion-specific features relevant to quality perception are extracted and used for quality estimation. Li et al. [78] propose a BIQA method based on the blur distortion. They first calculate the gradient image to characterize the blur distortion. Then, they divide the gradient image into blocks and extract the energy features of each block relevant to the blur distortion. Finally, the image quality is obtained by normalizing the moment energy. However, when image is distorted via unknown distortion channels, it becomes much more difficult to find specific features to measure image quality.

Recently, in order to assess the image quality without the prior knowledge of distortions, the non-distortion-specific BIQA methods have been developed. The natural scene statistics (NSS)-based methods are widely used to extract reliable features, which assume the natural images share certain statistics and the occurrence of distortions may change these statistics [14]–[16], [80]–[82]. In [14], [16], they aim to utilize NSS model, including the multivariate Gaussian (MVG) model [14] and the Generalized Gaussian distribution (GGD) model [16], to extract low-level image features for quality prediction. Although those methods have greatly improved the BIQA performance, there still exists a large gap between prediction scores and subjective scores. In order to further improve prediction performance, Wu et al. [15] use the multi-channel fused image features to simulate the hierarchical and trichromatic properties of the human vision. Then, the k-nearest-neighbor(KNN)-based model is used to evaluate image quality. Similarly, Ji et al. [80] assume that image quality is highly correlated with the degraded visual information. Therefore, they use the joint entropy of degraded features to assess image quality, which stimulates the visual information of the images. Instead of studying the quality-relevant image features, Wu et al. [81] focus on exploring efficient learning models. They propose a novel local learning method to improve the prediction performance, which is beneficial to the training of the complex and large data sets.

However, the obvious limitation of those BIQA methods is that the hand-crafted features may not be able to adequately represent complex image structures and distortions. Therefore, to improve prediction performance, attempts have been made to adopt deep BIQA methods, recently. The motivation is that the deep neural network (DNN) can automatically capture more deep features relevant to quality assessment and optimize these features to improve prediction performance by using back propagation method. Therefore, the DNN can be applied to various image quality assessment (IQA) methods [83], [84] and provides a very promising option for addressing the challenging BIQA task.

It is well known that deep learning techniques have achieved great success in solving various images recognition and object detection tasks [17]–[20]. The main reason is that it relies heavily on large-scale annotated data, like the image recognition oriented ImageNet [21] dataset. Unfortunately, for BIQA task, since there is a lack of sufficient ground truth labels IQA data for training, it is difficult to straightforwardly apply DNN to BIQA directly. This is because the DNN can lead to overfitting phenomenon, which means the trained model would have a perfect performance for training data but the performance is unreliable for unseen data. Therefore, researchers in the image quality community pay more attention to explore the useful DNN-based methods to solve this problem.

Previous surveys have also been summarized for BIQA methods, including classical methods [22]–[24] and DNN methods [25], [32]. However, the surveys of classical methods lack the analysis of the popular DNN methods [22]–[24]. And although some DNN methods are reviewed in [25], these methods can only be applied to the case where DNN input is the image patch. At present, there are still many novel DNN methods that have not been summarized [26]–[31]. In addition, a simple comparison of different DNN methods is represented in our previous work [32], but we have not made a comprehensive analysis and evaluation of various DNN methods, including the design strategy, network architecture, network complexity, advantages and disadvantages.

Therefore, in this paper, we intend to systematically analyze the various DNN methods, which aims to summarize the intrinsic relationship among various DNN methods. First, according to the different role of DNN, we divide the DNN methods into two categories, which could distinguish different DNN methods easily. One is the support vector regression (SVR)-based BIQA methods, which use DNN to extract deep feaures and SVR methods to predict image quality. The other is the DNN-based BIQA methods, which takes full advantage of back-propagated capability of DNN to optimize prediction accuracy. Moreover, we analyze the first type of DNN methods according to whether the input of DNN is low-level features or image/image patch data. Similarly, we analyze the second type of DNN methods according to the difference of DNN output. Fig. 2 shows the classification of different DNN methods, which aims to better understand different DNN methods easily. Finally, we summarize useful findings and discuss the challenges of DNN methods for BIQA. We hope that this study will be beneficial for the researchers to better understand this field.

Our contributions can be summarized as follows.

1)According to the different roles of DNN, we propose a new classification method, which could distinguish and improve understanding different DNN methods.

2)We analyze the DNN methods proposed in recent years, in terms of the contributions, the network architecture, the complexity, and the advantages and disadvantages. Especially, we also summarize many novel DNN methods that have not been discussed in previous literature surveys.

3)We systematically evaluate the prediction performance in difference DNN methods and obtain some interesting conclusions. Meanwhile, we also discuss some potential challenges and solutions for future research.

The rest of this paper is organized as follows. In Sec. II, we reviews the methods of SVR-based image quality prediction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

using deep features extracted by DNN. In Sec. III, we reviews the methods of DNN-based image quality prediction in detail and compare the implementations of these methods. The prediction performance and complexity of different DNN methods are analyzed in Sec. IV. In Sec. V, we provide some notable challenges of DNN-based BIQA methods. Conclusions are given in Sec. VI.

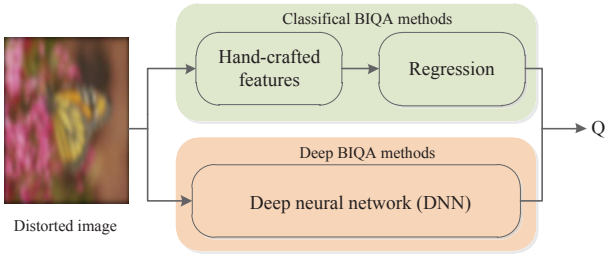

FIGURE 1: The flowchart of existing BIQA methods.

II. SVR-BASED IMAGE QUALITY PREDICTION USING DEEP FEATURES EXTRACTED BY DNN

Since the deep features from DNN can capture more useful information related to image distortions and human perceptions [25], the straightforward approach to employing DNN models is to extract discriminative deep features for various distorted images, and then evaluate the image quality using conventional SVR method. Recent work in the literature using DNN to extract deep features can be classified into two major schemes: 1) extracting from low-level features of image and 2) extracting from data of image/image patch. Figure 3 shows the flow diagram of these methods [33]–[35], [37]–[39].

A. DEEP FEATURES EXTRACTED FROM IMAGE LOW-LEVEL FEATURES

This kind of method aims to feed a large number of low-level image features relevant to quality perception into a DNN to evaluate image quality. Commonly, the low-level features are based on the NSS and other complementary features, which can accurately describe the structure features of distorted images. Then, these low-level features can be fed into the pre-trained DNN, including deep belief network (DBN) or stacked auto-encoder (SAE) network [33]–[35], to extract deep features. Especially, the unsupervised training method [36] is adopted to pre-train the DBN or SAE network. The goal is to overcome small IQA database problem and initialize each layer parameters of the pre-trained the DBN or SAE network. Afterwards, the parameters of entire network are fine-tuned with the labeled image features. Finally, the deep features extracted from the DBN or SAE model, along with the corresponding subjective scores are used to evaluate image quality by SVR method. Table 1 shows the details of these methods.

Tang et.al. [33] extract three types of low-level features, including NSS, texture, and blur/noise features. The NSS and texture features include the univariate and cross-scale histograms and statistics of complex wavelet transform of images (the real part, absolute value, and phase). These features aim to describe image global and local distortions. The blur/noise features include the patch PCA singularity [86], the two color model coefficient histograms [87], and the step edge based blur/noise estimation [88]. The blur/noise features can be added because these distortions are fundamental to various distortion types. Then, all of these low-level features are used to pre-train each layer of the DBN. And, the low-level features of IQA database with ground truth scores are used to fine-tune the entire DBN. Finally, a Gaussian process regression is used to obtain synthetic image quality score.

Ghadiyaram et al. further extend this work in [34] by combining DBN with SVR to predict authentically distorted images' quality. They adopt FRIQUEE method to extract low-level features of authentic images. FRIQUEE [77] first constructs several feature maps in multiple color spaces and transform domain, including luminance feature maps, LAB feature maps, and LMS feature maps. Then, the GGD, AGGD, and wrapped Cauchy models are used to fit feature maps and extract statistical features. Finally, these low-level features can be fed into a DBN model with extracted deep features and image quality scores are predicted by using SVR method.

In addition, Lv et al. [35] further improved the prediction accuracy and generalization ability. The authors select the multi-scale difference of Gaussian (DoG) features that are highly correlation with perceptual quality. This is because DoG is believed to simulate the neural processing procedure of how eye extracts details from images and convey them to the brain. Then, the SAE model is used to obtain deep features. Finally, these deep features are used to train an SVM regression model to predict image quality.

Compared with traditional BIQA methods, the major advantage is deep features extracted from low-level features is highly related to quality degradation. But the limitation is hand-crafted low-level features need to be carefully designed as the input to DNN, which does not take full advantage of DNN.

B. DEEP FEATURES EXTRACTED FROM IMAGE/IMAGE PATCHES

It is also observed that the deep features can be effectively mined by feeding data of image or image patches into the pre-trained DNN [37]–[39] for classification or recognition task, such as AlexNet [17], GoogleNet [18], RestNet [19], VGGNet [20]. Since the IQA is the human visual perception of the high-level semantics [40], the methods of image or image patches as DNN input can avoid the limitation of selecting low-level features to represent image high-level semantics accurately.

More specifically, some methods use image patches to extract deep features and these deep features derived from image patches are aggregated or pooled. Then, the predicted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

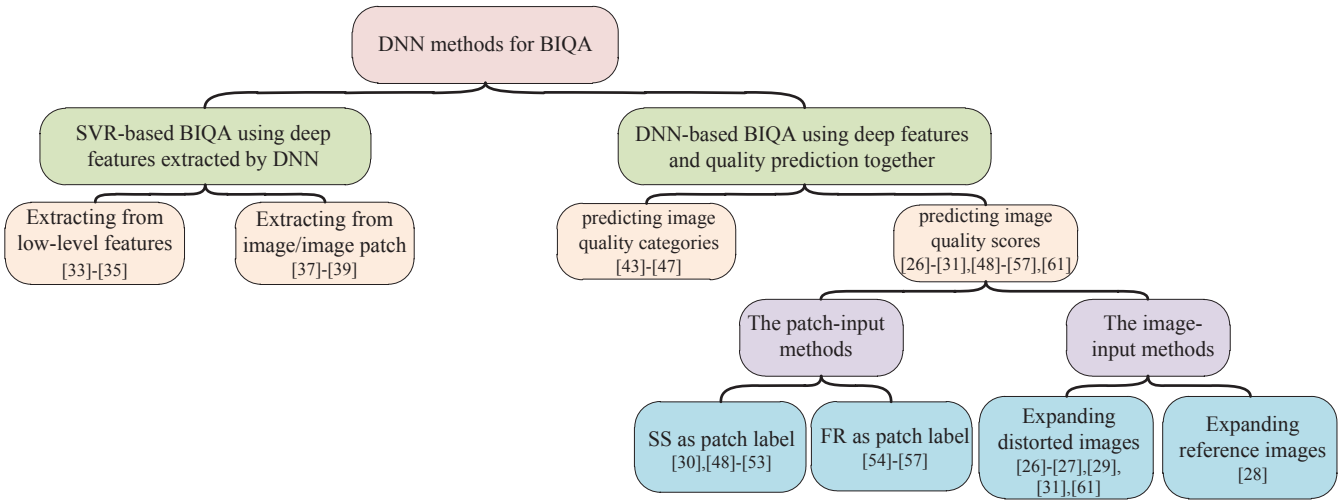

FIGURE 2: The classification of DNN methods for BIQA.

TABLE 1: The details of these methods [33]–[35]

Algorithms	Low-level features	DNN	Regression
[33]	NSS,texture,and blur/noise features	DBN	Gaussian process
[34]	Statistical features of GGD, AGGD, wrapped Cauchy models	DBN	SVR
[35]	multi-scale difference of Gaussian (DoG) features	SAE	SVR

quality of images is obtained by SVR method. In [37], the authors use multiple overlapping image patches as input to represent the whole image. They select the optimal layer of the pre-trained DNN model to extract deep features of each patch. Then, three kinds of statistical methods can be adopted to aggregate high-level semantic features of different patches. These aggregated features related to the whole image are fed into a linear regression model to predict image quality.

In addition, the deep features involving high-level semantic information of images are often used to evaluate image quality [38], [39], which is more consistent with human perception of images [41]. Sun et al. [38] proposed a BIQA framework, which is inspired by the human visual perception of image quality that involves the integrated analysis of global high-level semantics and local low-level characteristics. They use the first fully-connected (FC) layer of pre-trained AlexNet architecture to extract deep features, which aim to represent high-level semantic features associated with global image content. In addition to considering the high-level semantics, they also utilize the saliency detection and Gabor filters to perform local low-level features relevant to local image content. These features are combined to evaluate overall image quality by using SVR method. Similarly, Wu et.al [83] hypothesize that different levels of distortion generate individual degradations on hierarchical features. Therefore, they propose a BIQA framework based on hierarchical feature degradation. They first extract the low-level image features based on the orientation selectivity mechanism in the primary visual cortex, and then they use the last layer of the residual network (ResNet50) to extract deep features of visual content. Combining with the low-level image features and deep features, the image quality score is predicted by SVR methods. To further improve the prediction accuracy, Gao et. al [39] exploit multi-level deep feature fusion method to evaluate image quality. They assume that using only the last few layers' deep features may unduly generalize over local artifacts. Therefore, multi-level features representation compensates for local degradations. A DNN model formed by the pre-trained VGGNet is used to extracted image deep features over each layer. Afterwards, they utilize the SVR method to estimate the quality score from each layer's feature vector. The image quality is estimated by averaging layer-wise predicted score.

Considering that training a deep network is typically difficult for the small IQA database, these methods tackle the insufficient IQA database by extracting deep features from the pre-trained DNN model. Meanwhile, instead of selected low-level features as network input, the mehtods of deep features extracted from image or image patch data directly are more accurate. However, since the deep features extracted from the pre-trained DNN aims to deal with classification or recognition tasks, applying these features directly to our IQA task may not all be useful.

III. DNN-BASED BIQA USING DEEP FEATURES AND QUALITY PREDICTION TOGETHER

Instead of using DNN models to extract deep features related to quality degradation, this method directly uses the DNN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

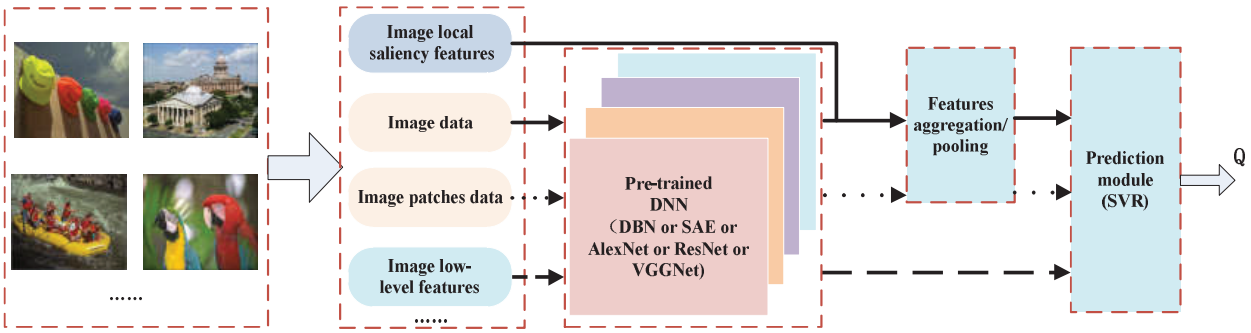

FIGURE 3: The flowchart of extracting deep features methods from DNN in [33]–[35], [37]–[39]

model to predict image quality. According to different evaluation metrics for quality prediction, there are two kinds of popular evaluation methods in recent years: predicting image quality categories and predicting image quality scores.

A. PREDICTING IMAGE QUALITY CATEGORIES

The DNN methods of predicting image quality catagories can be used to predict image quality categories, such as excellent, good, fair, poor or bad [42]. These labels have explicit semantic meanings in different quality ranges, so the category results can be directly used to describe the image quality. Meanwhile, the categorical quality assessment is a natural and viable way for human perception and can potentially reduce the randomness of the quality scores. Therefore, this kind of method treats BIQA as a classification problem to satisfy human visual behaviors. [43]–[47]. The general flowchart of these methods is shown in Figure 4.

Hou et al. [43] design deep network to classify images to five grades-excellent, good, fair, poor, or bad corresponding to the five point quality scale recommend by the International Telecommunication Union. The low-level features of NSS relevant to gray images can be extracted in the wavelet domain and fed into the DBN for layer-by-layer pre-training. Then, they recast image quality into five grades by using subjective method. Finally, they fine-tune the DBN to classify image grades by maximizing the probabilistic distribution. Further, considering not every region contributes to image quality perception, Hou et al. [44] also propose saliency-guided deep framework to improve prediction performance. First, they extract salient patches of natural image and adopt independent component analysis (ICA) method to learn basic filters. The same procedure can be applied to encoder salient patches of distortion image. The image-level features are a histogram that represents the frequency of learned ICA filters. Second, the DBN is pre-trained by layer-wise learning method and is fine-tuned by discriminative learning method, which makes the deep network can classify image grades.

The previous works pay attention to describe how to construct deep network but ignore to provide a clear understanding of why their framework performs so well. In [45], the authors not only propose a SAE method to classify image grades but also try to give a visualization explanation of how it works and why it works well. This is the first time to analyze and visualize deep network framework. Similar to the methods in [43], [44], they derive NSS-based features from shearlet-transformed RGB images and use the SAE model to classify seven quality grades that the train process is similar to DBN. In addition, they visualize the progression of training features to understand the DNN framework in the fine tuning stage.

The disadvantage of these methods is that the handcrafted features as network input cannot completely represent image distortions and contents. In order to overcome this problem, Bianco et al. [46] propose the end-to-end DNN framework to improve the prediction performance. They first pre-train AlexNet for classification task, which use 3.5 million images to pre-training from the ImageNet and Places databases. Then the pre-trained AlexNet is fine-tuned to classify the five image quality grades. Further, the prediction performance is better than the previous methods [43], [44].

In [47] a vector regression DNN model is proposed to obtain image quality grades. They divide image scores into five ordered intervals in response to five different grades. A belief score vector is computed by (1) to describe the probabilities of an image being assigned to different quality grades.

$$\vec{S} = \{s_1, s_2, s_3, s_4, s_5\} \quad s_k = y - u_k \quad k = 1, 2..., 5 \quad (1)$$

where $\vec{S}$ is a belief score vector, which collects five quality grade; s_k is the defined belief score to describe quality grade; y is the mean opinion score (MOS) of an image.

The DNN is trained to capture the associated belief score vector. It suggests that the smaller the value of $|s_k|$ is, the image quality is closer to the k-th grade. Finally, they propose an object pooling strategy to convert image quality grade into score, which fully takes into account the influence of the salient objects on image quality.

Although prediction grade methods are much more natural

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

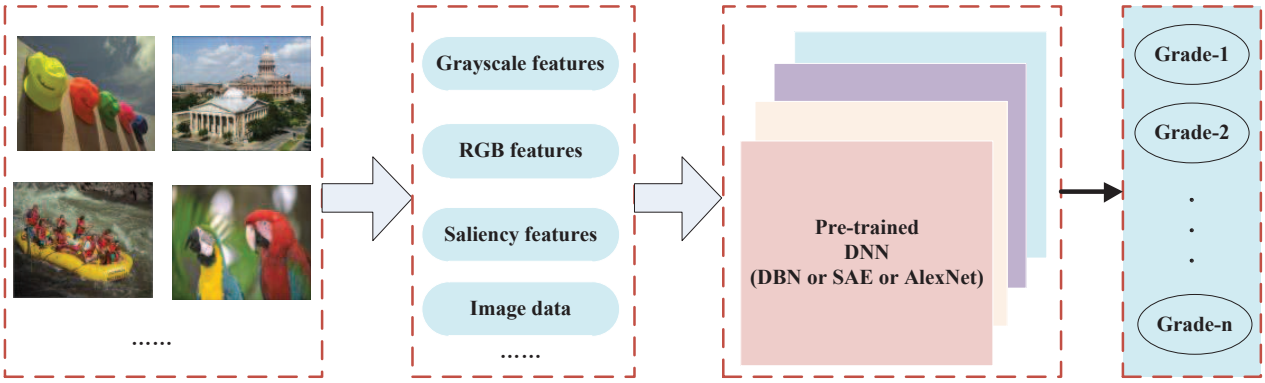

FIGURE 4: The flowchart of predicting image quality categories' methods in [43]–[46]

to evaluate image quality, the drawback is that different definitions of grades of subjective opinions can significantly impact the prediction performance of algorithms. Meanwhile, in order to make a fair comparison with other algorithms, the qualitative evaluations are converted into numerical scores by using different methods. Different conversion methods will also affect the final evaluation performance.

B. PREDICTING IMAGE QUALITY SCORES

The methods of predicting image quality scores are the most popular for BIQA. The characteristic of this method is purely data-driven and allows for end-to-end optimization of feature extraction and regression. It means that these DNN methods can be used to predict image quality scores, such as DMOS=72.34, DMOS=25.2. This gives a specific scalar as a score to measure image quality. Especially, most of DNN methods adopt this approach to predict image quality, because many of IQA databases use scalar scores to describe image quality. Therefore, in order to keep the predicted results in consistent with the IQA databases, this kind of method can be treated as a regression problem. Although previous work has summarized this method [25], it only introduce the methods using image patch as DNN input and some novel DNN methods that have been appeared recently are not analyzed [26]–[31], [53], [54], [57]. Thus, we will systematically summarize and analyze the existing methods. According to different input in DNN, we propose a classification method: the patch-input methods and the image-input methods.

1) The patch-input methods

The performance of DNN heavily depends on the number of training data. However, the currently available IQA databases are much smaller compared to the classification or recognition tasks [17], [18]. Moreover, obtaining large-scale reliable human subjective labels is very difficult. To expand the training database, the patch-input method aims to divide image into multiple patches as DNN input to increase training samples.

There are many methods based on image patches as DNN input. According to the different labels of training patches, we discuss these methods in two ways. One is to use the image subjective score (SS) as image patch label [30], [48]–[53]. The other is to use FR as image patch label [54]–[57].

a: SS as image patch label methods

In [48], this is the earliest method that integrates feature learning and patch quality prediction into an end-to-end network. They divide gray images into 32×32 patches. Each image patch with image subjective score as input is used to train DNN, which consists of 1 convolutional (C), 2 pooling (P) and 3 full-connected (FC) layers. The image quality is estimated by the average score of all image patches. Nevertheless, the problem is that they ignore that the visual quality of different local regions is often different and humans tend to concentrate on the regions of saliency when evaluating an image. Therefore, the salient patches of images can be considered to predict image quality in the following methods [49]–[51].

In [49], the authors design a seven-layer DNN architecture to capture patch-level quality prediction focusing on color images. They then perform the saliency detection with free energy based neural theory to obtain image saliency map [58]. After that, they define the weights of image patches by the corresponding saliency map. The final image quality score is yielded with the weighted average of each image patch. To further improve prediction performance, in [50], [51], they consider only the salient patches to evaluate image quality score. First, they also split the image into patches and use typical saliency detection methods to obtain image saliency map. Further, they assign a threshold to remove non-salient patches. The remaining salient patches are reweighted into the range of [0, 1]. The whole image quality score is calculated by the weighted average over salient patches. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

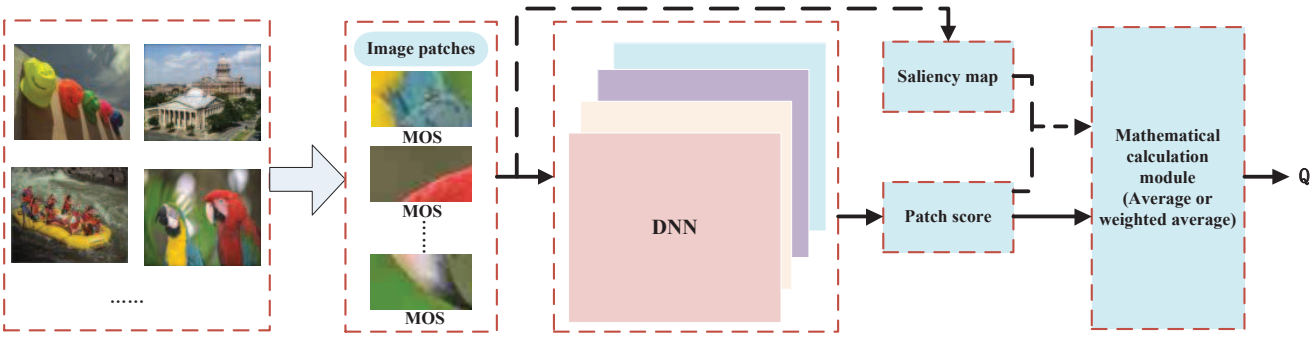

FIGURE 5: The general flowchart of SS as image patch label methods in [49]–[51]

general flowchart is shown in Figure 5.

However, the previous weights of saliency maps are set artificially, which is inaccurate to image quality. Some methods study the use of end-to-end DNN to simultaneously obtain patches' scores and corresponding weights. The weights obtained by DNN learning method more accurately respond to the image perception. In [52], the distorted image patches can be fed into DNN, which consists of 9 C layers, 5 P layers for feature extraction and 2 FC layers for regression. The role of first FC layer of DNN architecture is used to learn patches' weights and the second FC layer is used to learning patches' scores. The image quality score is calculated by weighting average of all patches' scores. Compared with the models employing simple average pooling or artificial setting weight pooling, this method improves prediction accuracy and has well generalization ability. Similarly, in [53], they also divide image into 100 image patches and fed them into the DNN to obtain patch score and weight. Considering the relationship between image contents and patches' weights, the global regression layer is used to optimize image prediction score.

In addition, in order to learn the complicated relationship between visual appearance and the perceived quality, Yan et al. propose a novel two-stream DNN architecture, which takes the raw image and the gradient image as input visa two sub-networks [30]. The motivation of this design is to integrate input information from different domains to represent the quality of distorted images. Each image is divided into different patches as the inputs of the image stream sub-network. Each of the sub-network consists of ten layers to extract image features. Especially, the region-based full convolutional layer is used to handle the locally non-uniform distortions of images. The gradient stream sub-network is similar to image stream and the input is gradient patches. Then, a concatenate layer is used to fuse features from the two streams and the followed three FC layers are used to predict patch quality. Finally, the quality score of the whole image is calculated by averaging all patches' scores. The overall framework of the algorithm is presented in Figure 6.

Table 2 compares the implementation of reported patch-input algorithms, which the path label is the ground truth score. It is worth nothing that C, P and F mean convolutional layer, pooling layer and full-connected layer, respectively. w_i means the weight of the i-th patch. M means the number of all patches of an image. K means the number of salient patches of an image. q_i is the prediction patch score from DNN model. In table 2, we find that because of the increase of training samples, the patch-input algorithms can design a deeper network to evaluate image quality score. Meanwhile, these methods mainly pay attention to the effect of salient patches on image quality. However, the labeling of image patches with the whole image subjective score is problematic, because the ground truth score for each patch does not exist. In addition, the whole image quality score is calculated by the sample mathematical method, which may affect the accuracy of image quality prediction.

b: FR as image patch label methods

To overcome the problem of inaccurate patch label, the strategy that FR methods are used to calculated proxy score of image patch has been studied [54]–[57]. Figure 7 shows the flowchart of these methods.

In [54], it is a novel completely blind DNN methods. By taking the large scale of image patches as the training set, the authors design a feature fusion DNN in different layers and use FSIM as the label to train DNN architecture. The DNN consists of 6 C layers, 1 P layer, 2 sum (SU) layers and 2 FC layers. The role of the sum layer is to fuse different layer features to prevent gradient vanishing [19]. Especially, the training patch label is calculated by using the FR method, which is an accurate method to calculate patch label without subjective scores.

In [55], J. Kim et al. propose a two-stage DNN-based to evaluate image quality. The patch quality score generated by FSIM method are used as proxy patch label in the first stage of training. In the second stage, the feature vectors obtained from image patches are aggregated using statistical moments and then a global regression layer is used to predict image quality score. Rather than using complex DNN to produce proxy scores, the same authors develop a novel DNN, which aims to regress into objective error maps [56]. In the first stage, the objective error maps are used as proxy regression targets to train DNN, which is calculated by the absolute

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

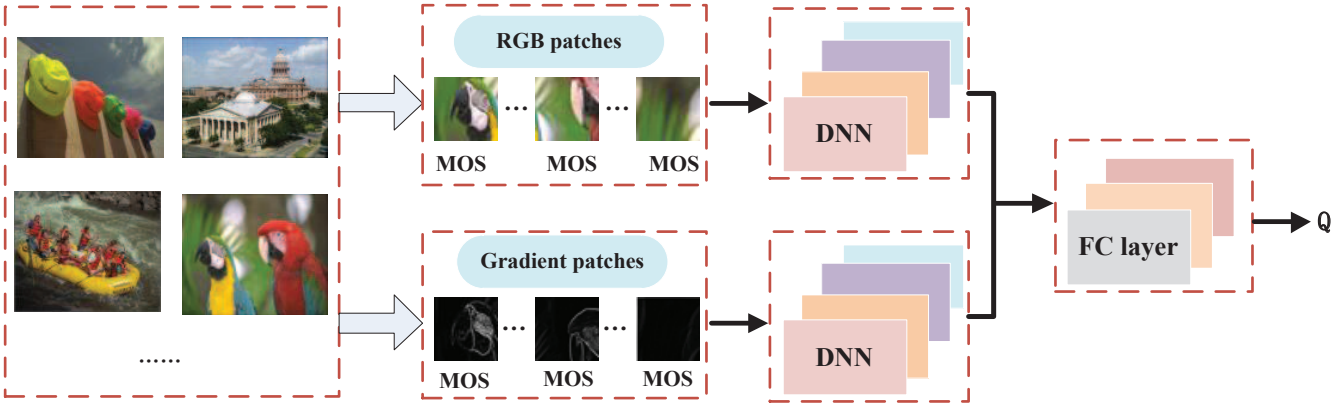

FIGURE 6: The overall framework in [30]

TABLE 2: The comparison of DNN methods by using SS as patch label in [30], [48]–[52]

Algorithms	Layer depth	Path size	Saliency methods	Assigned weight	Selected patch	Image score
[48]	1C,2P,3F	32×32	N/A	N/A	All patches of an image	$Q = \frac{\sum_{i=1}^{M} w_i \times q_i}{\sum_{i=1}^{M} w_i}$
[49]	2C,2P,3F	32×32	Free energy	$w_i = \sum_{j=1}^{N \times N} S(j)$	Saliency patches	$Q = \frac{\sum_{i=1}^{K} w_i \times q_i}{\sum_{i=1}^{M} w_i}$
[50]	2C,4P,3F	32×32	Fast SM	$w_i = \sum_{j=1}^{N \times N} S(j)$	Saliency patches	$Q = \frac{\sum_{i=1}^{K} w_i \times q_i}{\sum_{i=1}^{M} w_i}$
[51]	10C,4P,3F	32×32	SDSR	$w_i = \sum_{j=1}^{N \times N} S(j)$	Saliency patches	$Q = \frac{\sum_{i=1}^{K} w_i \times q_i}{\sum_{i=1}^{M} w_i}$
[52]	9C,4P,2F	32×32	N/A	DNN learning	All patches of an image	$Q = \frac{\sum_{i=1}^{M} w_i \times q_i}{\sum_{i=1}^{M} w_i}$
[30]	10C,8P,3F	32×32	N/A	N/A	All patches of an image	$Q = \frac{\sum_{i=1}^{M} w_i \times q_i}{\sum_{i=1}^{M} w_i}$

difference between the reference image patch and distortion image patch. In the second stage, the extracted feature maps from DNN are fed into the global average pooling layer, then regress onto ground-truth scores by using two fully connected layers. The prediction accuracy is competitive with the state-of-the-art methods.

To further improve prediction performance, Pan et al. propose a novel framework for BIQA, which consists of a generative quality map network and a quality pooling network [57]. They employ MDSI [59] to generate patches' quality maps as labels and select U-Net [60] as a base of generative network to train image patch quality map. The output quality maps are fed directly into the pooling network to regress patches' scores. Finally, the final score of the whole image is obtained by using the average of all image patches' scores.

Table 3 compares these algorithms to obtain patch label by using the FR methods. Compared with the methods of subjective score as patch label, the FR metrics are used as intermediate local targets for each image patch, which reduce the error of using the whole image subjective score as patch label. In addition, instead of the simple mathematical calculation to obtain image quality score, the global optimization method is more accurate for DNN. Whereas, the disadvantage of using FR methods as patch label is that it is very hard to obtain reference images in many practical applications for the FR metrics.

2) The image-input methods

Rather than using image patches as the input, the image-input methods aim to train a prediction model by using the whole image and its associated ground truth, which can effectively overcome the difficulty of being able to obtain the ground truth of image patches. However, there has been limited effort towards end-to-end optimized BIQA using DNN, primarily due to the lack of sufficient ground truth labels of images.

Recently, the image-input methods are developed [26]–[29], [31], [61]. The novelty is that, despite a lack of image databases, the DNN based on image as input can also evaluate image quality very well. This is because the image expansion techniques are used to solve insufficient IQA database. According to the different extended objects, we classify these methods into two sub-categories: expanding distorted images and expanding reference images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

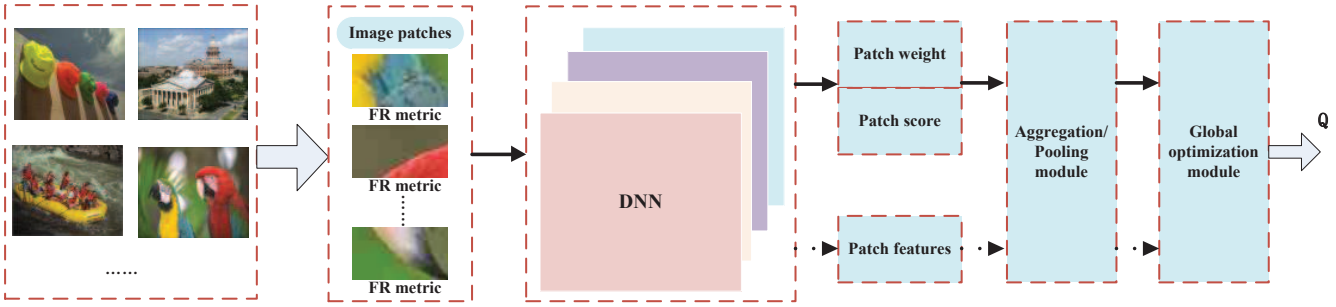

FIGURE 7: The flowchart of FR as image patch label methods in [54]–[57]

TABLE 3: The comparison of DNN methods by using FR as patch label

Algorithms	Layer depth	Path size	The type of label	FR metrics	Image score
[54]	6C,1P,2SU,2F	32×32	A score value	FSIM	Global optimization
[55]	2C,2P,6F	32×32	A score value	FSIM	Average of patches' scores
[56]	9C,2P,3F	112×112	Error map	Absolute difference	Global optimization
[57]	13C,2P,2F	114×114	Quality map	MDSI	Average of patches' scores

a: Expanding distorted images' methods

For expanding distorted images' methods, two expanded ways are shown: large databases, such as the ImageNet [21], Places2 [62], and the artificial generation images [26]–[29]. The DNN then is trained by the transfer learning method [63]. This is a common way to overcome the small database task.

When the distorted images come from the large database, these distorted images can be used to pre-train a DNN. Then, the small IQA database is used to fine-tune the pre-trained DNN to evaluate image quality score. In [61], Li et al utilize Network in Network (NIN) [64] and transfer learning technique to deal with BIQA problem. The first step is that the NIN is pre-trained for the classification task on the large-scale ImageNet database. Through this pre-training process, the good initial weights can be obtained, which is much better than randomly initialized weights. In the second step, they modify the pre-trained NIN architecture, which the final layer is replaced by regression layers. In the third step, only the small IQA images with ground truth scores are used to fine-tune the pre-trained NIN. However, for synthetic IQA database, such as LIVE [65], TID2013 [66], CSIQ [67], LIVE multiply distorted (MD) [68], the prediction performance is not accurate. This is because the pre-trained NIN learns the features of authentic distortions of the ImageNet database, which is different from synthetic distortions.

In [31], they assume that various kinds of distortions exist in different IQA databases, which requires different level features to predict visual quality. Therefore, they propose a DNN model using multiple levels of features simultaneously to achieve a consistent performance over different IQA databases. The ResNet-50 [19] model which is pre-trained on the ImageNet database is adopted as baseline. In the fine-tuning stage, they divided all ResNet blocks into four groups and extract each group's features. Then, they define an encoder layer to unify the feature size from different levels. Finally, these multiple levels of features are combined and fed into the FC layer to evaluate image quality score. This method shows the state-of-the-art accuracy on different IQA databases.

Besides, the artificial generation method [26], [27], [29] can be used to construct the large-scale pre-training distortion images, which is similar to the IQA database. It is far from realistic to carry out a full subjective test to obtain a MOS/Difference MOS (DMOS) for each image. Whereas, the challenge of this method is how to obtain the ground truth labels of generated images in the pre-training stage.

To overcome this problem, the motivation of Rank [26] is to design a new strategy to generate the large-scale distortion images without laborious human labeling. According to the rule that the image quality decreases with the increase of the distortion levels, they synthetically generate the ranked image pairs with five different distortion levels from Waterloo Exploration database [69]. The Waterloo Exploration database contains 4744 pristine images and covering various image contents. Especially, the generated distortion image pairs are similar to the IQA database. In the LIVE database, they exclude fast fading (FF) distortion type and generate the remaining four distortion types: JPEG compression (JPEG), JPEG2000 compression (JPEG2000), additive while Gaussian noise (WN), Gaussian blur (GB). In the TID2013 database, they generate 17 out of a total of 24 distortion types. Moreover, we do know for any pair of images which is of higher quality. Then, using the pairs of the ranked images, we pre-train a Siamese network [70] to learn image distortion levels by using the proposed Siamese back-propagation technique. Finally, they fine-tune a branch of Siamese network to predict image score, which aims to transfer image distortion levels to quality scores. Figure 8 shows the flowchart of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

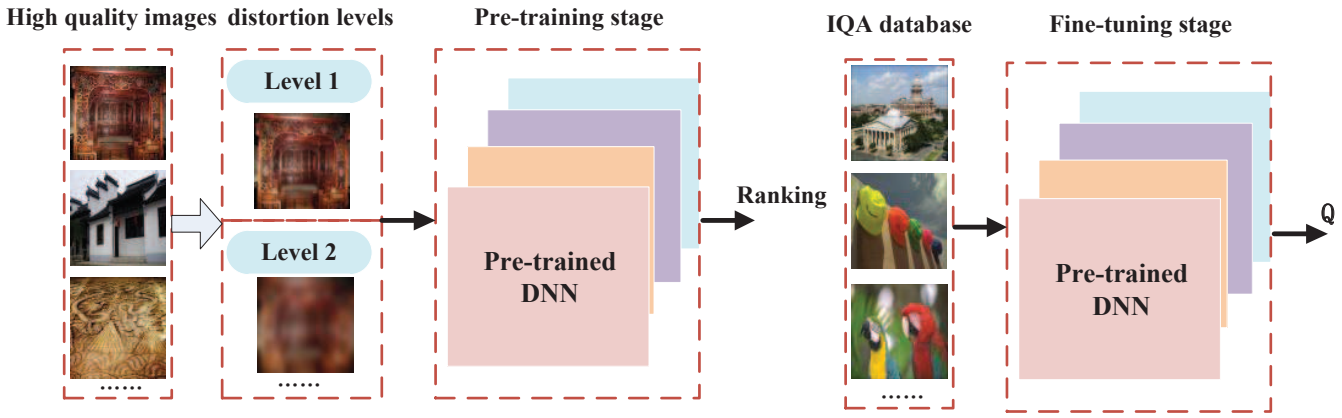

FIGURE 8: The flowchart of Rank method in [26]

Rank method. Compared with existing BIQA methods, the prediction performance is the best in LIVE database and even outperforms the state-of-the-art in FR methods.

However, the limitation of the Rank method is that it can only simulate distortion images in artificially synthetic IQA database, but it is difficult to apply this method to authentic IQA database. This is because we cannot know the priori information of authentic distortion images. Therefore, to improve performance of different IQA databases, Zhang et al. design an end-to-end DB-CNN solution for BIQA that works for both synthetically and authentically distorted images [27]. First, they describe the generation process of the large-scale database in the pre-training step. They use two large-scale databases: Waterloo Exploration database and PASCAL VOC 2012 [71] to generate distorted images. Considering the distortion types of the synthetic IQA databases, they produce nine distortion types related to the LIVE, TID2013, CSIQ and LIVEMD databases, i.e., JPEG, JPEG2000, WN, GB, pink noise, contrast stretching, image quantization with color dithering (ICQD), over-exposure and under-exposure. Especially, the first six distortion types cover the entire CSIQ database. Meanwhile, they synthesize distorted images with five distortion levels except for over-exposure and under-exposure, for which only two levels are generated. In summary, the pre-training database contains 852891 distorted images. The ground truth label is presented as a 39-class indicator vector to encode underlying distortion types at the specific distortion level. The dimension of ground truth vector comes from the fact that there are seven distortion types with five levels and two distortion types with two levels.

Then, they design the architecture of the S-CNN for synthetically distorted images, which consists of 9 C layers, 1 P, 3 FC layers and a softmax (S) layer. It aims to classify the probability of each distortion type at the specific degradation level. Considering this DNN model is not beneficial for authentic IQA databases, they select the pre-trained VGG-16 network for the classification task on ImageNet as another branch to extract relevant features for authentically distorted images. This is because the distortions in ImageNet occur as a natural consequence of photography rather than simulations. Finally, in the fine-tuning step, they tailor the pre-trained S-CNN and VGG-16 and introduce bilinear pooling module to combine the S-CNN for synthetic distortions and VGG-16 for authentic distortions into a single model, which aims to discriminate synthetic or authentic distortions. The FC layer follows the bilinear pooling layer to predict image quality score. The flowchart of DB-CNN can be shown in Figure 9.

A closely related work to DB-CNN [27] is MEON [29], a cascaded multi-task DNN framework for BIQA. This method also pays attention to the influence of distortion types and levels on quality degradation. Figure 10 shows the flowchart of MEON method. The subtask I aims to pre-train a distortion type identification network, for which large-scale training samples are readily available. They select 840 high-resolution natural images to generate C distortion types' images and each distortion type images has five distortion levels. The ground truth label is a C-dimensional vector to encode distortion types. This network consists of 4 C layers, 4 P layers, 2 FC layers and 1 S layer. Especially, they choose biologically inspired generalized divisive normalization (GDN) instead of rectified linear unit as the activation function of C layers and FC layers. The sub-task II network appends two FC layers after the shared DNN architecture from sub-task I. Then, they define a fusion layer (FS) that combines the distortion types' features from sub-task I and the distortion levels' features from sub-task II to yield an overall quality score.

Table 4 summarizes the expanding distorted images' methods. LM means the learn method, GT means the ground truth of generation images and NGI means the number of generation image. We clearly see that the transfer learning method is used to overcome the small IQA databases. The pre-training DNN aims to resolve the classification problem, because the ground truth labels can be easily known instead of humans' subjective judgment. Especially, the depth of network is proportional to the number of pre-trained samples. Moreover, in order to deal with authentic images, they add the sub-network to meet the prediction of authentic IQA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

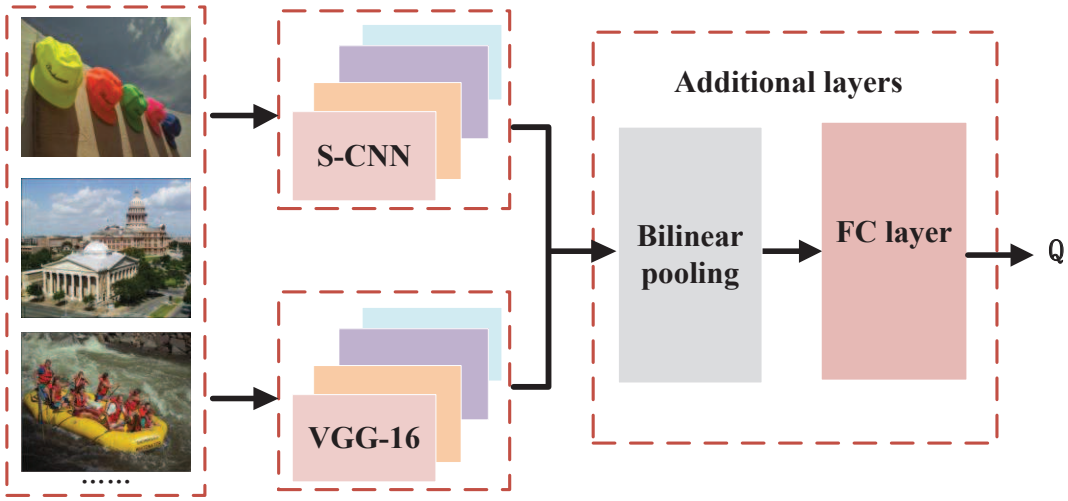

FIGURE 9: The flowchart of DB-CNN method in [27]

TABLE 4: The comparison of the image-input DNN methods

Algorithms	DNN type	Layer depth	Image size	LM	GT	NGI
[26]	VGG-16	13C,5P,3F	224×224	Transfer learning	Distortion levels	94880
[27]	Two sub-networks	22C,5P,1S,4F;	224×224	Transfer learning	Distortion types and levels	852891
[29]	Two sub-networks	8C,8P,1S,1FS,4F;	224×224	Transfer learning	Distortion types and levels	25200

database.

b: Expanding reference images' methods

This is a novel topic to use generative adversarial network (GAN) to augment images. Since the distortion images and corresponding non-distortion reference images are typically absent in IQA databases, it leads to the prediction performance of image quality being not accurate. Thus, the HIQA method [28] aims to address this problem by combining the GAN and the GAN-guided quality regression (R) networks. The Fig.11 shows the flowchart of the GAN method. First, the quality-aware generative (G) network can be used to overcome the absence of reference image, which aims to generate the hallucinated reference image I_h conditioned on the distorted image I_d. In order to reduce the difference between the hallucinated image and the corresponding reference image, the loss function of G network can be designed by using the pixel-wise error and the perception-wise difference. Second, they propose a IQA-Discriminator (D) network to adjust the loss of G to produce high perceptual outputs, even when G fails to generate hallucination images, the predicted scores of R network should still be reasonable value. Finally, the distorted images and their discrepancy maps between hallucinated images and its corresponding distortion images are fed into the R network and the high-level features fusion scheme is adopted to optimize R network. Especially, the training strategy is set. The GAN network is trained to generate a large number of the hallucinated images, which is similar to the reference images in IQA database. And then, the R network is trained to predict image quality score. In GAN network, the D network is first trained to distinguish the fake reference images from the reference images of the IQA database. Then, the G network is trained to generate images, which is similar to the real reference images in the IQA database. Finally, the image quality score can be predicted by optimizing the loss of the R network.

IV. THE PERFORMANCE OF DIFFERENT DNN METHODS

A. DESCRIPTION OF PUBLIC DATABASES AND EVALUATION METRICS

The choice of a database for training is important for deep-learning-based models, since their performance highly depends on the size of the training set. We briefly describe several popular public databases for BIQA, including LIVE [65], TID2013 [66], CSIQ [67], LIVE MD [68], LIVE In the Wild Image Quality Challenge Database (LIVEC) [72].

1) The LIVE database [65] includes 29 reference images and 779 distorted images degraded by five types of distortions (JPEG, JP2K, WN, GB, Rayleigh fast-fading channel distortion (FF)). Subjective quality scores are provided in the form of difference mean opinion score (DMOS) ranging from 0 to 100, where a lower score indicates better image quality.

2) The TID2013 database [66] contains the largest number of distorted images. It consists of 25 reference images and 3000 distorted images with 24 different distortion types at five levels of degradation. The database also provides the MOS, ranging from 0 to 9. A higher value of MOS indicates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

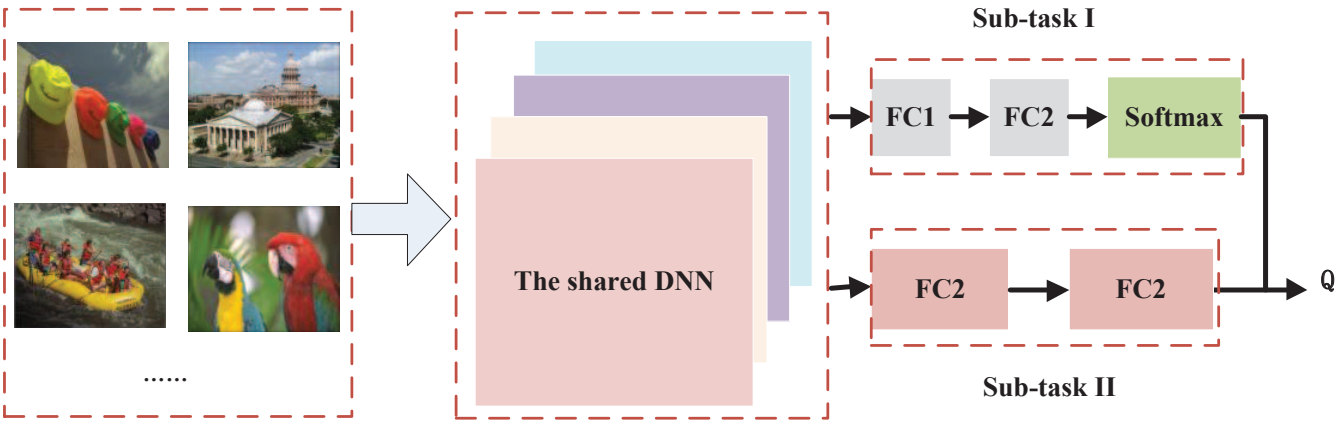

FIGURE 10: The flowchart of MEON method in [29]

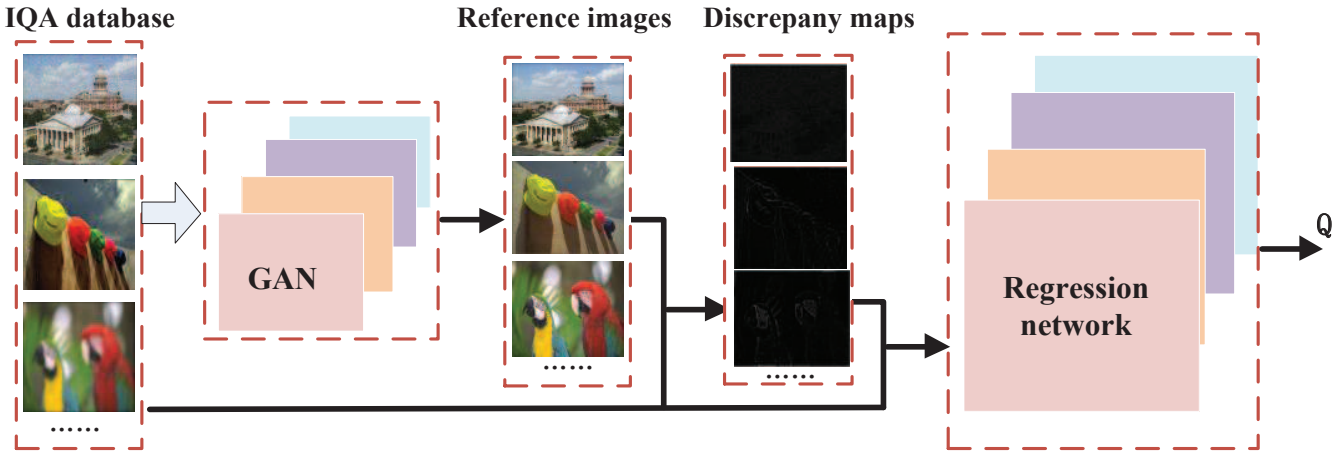

FIGURE 11: The flowchart of the HIQA method in [28]

higher quality. The distortion types include a range of noise, compression, and transmission artifacts.

3) The CSIQ database [67] consists of 30 reference images and 866 distorted images corrupted by six types of distortions: JPEG, JP2K, WN, GB, pink Gaussian noise and global contrast decrements. Each image is distorted by five different distortion levels. Subjective quality scores are provided in the form of DMOS ranging from 0 to 1.

4) The LIVE MD database [68] was the first to include multiple distorted images. Images are distorted by two types of distortions in two combinations: simulated GB followed by JPEG and GB followed by additive WN. It contains 15 references and 450 distorted images, and the DMOS of each distorted image is provided, ranging from 0 to 100.

5) The LIVE In the Wild Image Quality Challenge Database (LIVEC) [72] comprises 1162 images, which are captured using modern mobile devices and contain diverse authentic image distortions. In addition, no undistorted reference images are available in LIVEC. Subjective scores are obtained in the form of MOS in an online crowdsourcing platform. MOS values lie in the range [0, 100]. The summary of the above databases is shown in Table 5. Note that Ref means the number of reference images. Dist means the number of distorted images. DT means the number of distortion types. SST and SR mean subjective score's type and range.

TABLE 5: Comparison of different IQA databases

Database	Ref.	Dist.	DT	SST	SR
LIVE	29	779	5	DMOS	[0,100]
TID2013	30	3000	24	MOS	[0,9]
CSIQ	25	866	6	DMOS	[0,1]
LIVE MD	15	450	2	DMOS	[0,100]
LIVEC	N/A	1162	Numerous	MOS	[0,100]

Two commonly used metrics [73], Spearman Rank-Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) are used for performance evaluation. These metrics are to measure the correlation between a set of estimated visual quality scores Q_{est} and a set of human subjective quality scores Q_{sub}, as:

$$SROCC(Q_{est}, Q_{sub}) = 1 - \frac{6\sum d_i}{m(m^2-1)} \tag{2}$$

$$PLCC(Q_{est}, Q_{sub}) = \frac{cov(Q_{sub}, Q_{est})}{\sigma(Q_{sub})\sigma(Q_{est})} \tag{3}$$

where m is the number of images in the evaluation database; d_i is the rank difference of i th evaluation sample in Q_{est} and Q_{sub}; $cov(.)$ represents the covariance between Q_{est} and Q_{sub}; $\sigma(.)$ represents the standard deviation. The PLCC measures the prediction accuracy and the SROCC measures the prediction monotonicity. For both correlation metrics a value close to 1 indicates high performance of a specific quality measure.

B. PERFORMANCE COMPARISON ON INDIVIDUAL DATABASE

We compare the performance of a number of state-of-the-art BIQA and FR-IQA methods, including: FR-IQA methods (PSNR, SSIM [3], FSIMc [74], DeepQA [76]) and classic BIQA methods (BRISQUE [75], BWS [16], CORNIA [12], GMLOG [13] and IL-NIQE [14]), current leading various BIQA methods based on DNN (MGDNN [35], FRIQUEE [34], GLCP [38], BLNDER [39], DLIQA [43], SESANIN [45], VPOR [47], CNN [48], Pre-SM [50], VIDGIQA [53], DIQaM [52], TSCN [30], BIECON [55], DIQA [56], BP-SQM [57], MFIQA [31], RANK [26], DB-CNN [27], MEON [29], HIQA [28]).

For the classic BIQA methods and FR-IQA methods, we conducted experiments by utilizing the respective codes released by the authors. It is, however, difficult to reproduce the BIQA methods based on DNN. We therefore first adopted the results reported in the respective literature. Especially, for the cases where experimental results are not given, we use the released codes to conduct experiments and generate results, such as CNN, DIQaM, BIECON, RANK.

As shown in table 6, the SROCC and PLCC values are reported to various methods. The best three performances among the BIQA methods are shown in bold. The weighted average of the SROCC and PLCC over the five databases is shown in the last column of table 6. The weight of each database is proportional to the number of distorted images in the database. Especially, in table 6, NR1 means the classic BIQA methods. NR2 and NR3 mean the extracting deep features from low-level features and image/image patch methods, respectively. NR4 means the prediction grades' methods. NR5 means the SS as patch label's methods and NR6 means the FR as patch label's methods. NR7 and NR8 mean the expanding distorted images' methods and the expanding reference images' method, respectively.

We can see that the DNN methods generally perform better than the classic BIQA methods. The fundamental difference between DNN methods and classic BIQA methods is that, rather than using hand-crafted features and shallow regression for classic BIQA, DNN methods search for highly optimized features automatically and can significantly reduce prediction errors by the deep network. Meanwhile, we also show the RMSE performance in table 7. It can be clearly seen the RMSE performance of the DNN methods is better than the classical methods in LIVE database. In other IQA databases, the DNN methods are better than the classical BWS method. This is because the DNN methods can learn image deep features related to perception and use the back propagation method to train the deep network. Therefore, it is why the DNN methods have been developed rapidly to improve IQA performance in recent years. In addition, DNN methods are highly competitive with the FR methods. However, DNN methods do not use any prior information of reference for image quality assessment.

We compare the extracting deep features methods from DNN models [34], [35], [38], [39]. Although some methods do not give all the experimental results in the five databases, we clearly see that the methods of directly extracting from data of image/image patch [38], [39] are better than the methods of extracting image low-level features [34], [35]. The main reason is that the selected low-features are limited and cannot adequately describe the image distortions and contents. However, compared with other end-to-end DNN models, these methods are simple by using shallow regression method.

Compared with the methods of predicting quality grades [43], [45], [47]. The VPOR method significantly outperforms the DLIQ and the SESANIN methods in LIVE database. First, the image grade labels, which are defined in a belief score vector method, are more accurate than the subjective grades in DLIQ and the SESANIN methods. Second, when converting the image quality grade to the image score, the V-POR method take into account the influence of object saliency on image quality. It makes the prediction performance is better than the DLIQ and SESANIN methods. Therefore, we find that although qualitative classification methods are much natural to human visual behaviors, the classification of grades and the strategy of converting image score will affect the final prediction performance.

For the patch-input methods, there is a competition between SS as image patch label methods [30], [48], [50], [52], [53] and FR as image patch label methods [55]–[57]. When only the image subjective score is used to obtain image patch label, the prediction accuracy is inferior to the methods of using FR as patch label. It is clearly see that the BIECON, DIQA and BPSQM are all better than CNN. This is because FR method considers the visual sensitivity of the different image patch, so that the obtained patch label is more accurate than the whole image subjective score as label. However, after adding the saliency of the image patch, the subjective score methods is highly competitive with the FR methods to obtain image patch label. This is easy to understand because the differences can be highlighted after considering salient image patches. Whereas, although the FR methods and the image patches' saliency methods can approximately obtain the quality of different image patches, the obtain labels are not the real ground truth of image patches, because the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

ground truth quality of each patch does not exist.

For image-input methods, we clearly see that in the synthetic IQA databases, the methods of expanding distorted images [26]–[28] are more benefit than that of directly using large database methods [31]. In the LIVE database, the RANK, DB-CNN methods perform superior to the MFIQA, because artificial generation method can simulate images with similar distortion types and levels in synthetic IQA database. Hence, the DNN can roughly learn the features of similar distortion images with IQA database in the pre-training stage. On the contrary, in the LIVEC database, MFIQA method is better than RANK, because the pre-trained DNN in the large database learn the real distortion features. However, due to the limitation of synthetic distortion images, it cannot meet the needs of various databases, which leads to poor generalization ability in different databases. In order to overcome this problem, the DB-CNN method design two sub-networks that can satisfy both synthetic and authentic distortion, thus improving the prediction accuracy. In addition, the expanding distorted images' methods compete with the expanding reference images. However, it is worth noting that the popular GAN method is first used to solve insufficient IQA database problem.

C. PERFORMANCE ON CROSS-DATABASE

It is expected that a robust BIQA model that has learned on one image quality database should be able to accurately assess the quality of images in other databases. Therefore, in table 8, we compare the results of generalizability of the classic BIQA methods and DNN methods only in the synthetic distortion databases. But we do not consider train the DNN model on the authentic image distortion database (LIVEC). On the one hand, this is because some DNN methods need to use the reference images or simulated distortions method to train DNN model, such as DIQA [56], RANK [26], while the LIVEC is the authentic image distortions without the reference images or prior distortion types. On the other hand, because of the largely difference between synthetic and authentic images, many DNN methods do not discuss cross dataset test between synthetic and authentic datasets. Therefore, the compared BIQA methods are trained using all the images from one synthetic database, and then tested on another database. In the CSIQ and TID2013 databases, four overlapping distortion types (WN, GB, JPEG, JP2K) are used.

In table 8, it can be seen that the DNN method is the best performance when LIVE database is trained and other subset databases are tested. The MFIQA and DIQA obtain the better performance than other methods when CSIQ subset and TID2013 subset are trained, respectively. Therefore, the generalization ability of the end-to-end DNN methods is generally better than the classic BIQA methods and the extracted deep features' BIQA methods. This is because the end-to-end methods can use images/image patches data to learn deep features and reduce the prediction errors by back propagation method. However, the classic BIQA methods are limited in extracting hand-crafted features, which cannot completely represent the image structures and distortions. Meanwhile, the prediction performance of shallow regression, such as SVR, is not as good as that of deep regression network. Similarly, although the extracted deep features methods can further extract the deep features from the limited low-level features, the shallow regression restricts the generalization ability.

Furthermore, in DNN methods, the generalizability of the patch-input methods [48],[52],[56] is better than the image-input methods [26], [31]. The main reason is the patch-input methods use the images of IQA database to expand training samples to train DNN network, but the image-input methods expand the IQA database by using exterior images. These exterior images can be fitted as IQA images to expand IQA database. Because the difference between the fitted images and IQA images, it reduce the generalization ability of the DNN model.

D. THE COMPLEXITY OF DIFFERENT DNN METHODS

We calculate the complexity of different DNN methods as shown in table 9, including CNN, DIQaM, BIECON, RANK, DB-CNN. Especially, WPs and BPs mean the weight parameters and basis parameters, respectively. ATPs means the total parameters of the DNN. CTs means the parameters of all C layers and FTs means the parameters of all F layers.Since C and F layers are used to update network parameters, the complexity of algorithm is closely related to the C and F layers' parameters. In table 9, we clearly see that the complexity of CNN is lower than the DIQaM, BIECON, RANK, DB-CNN, because the number of layers of the DNN is smaller than that of other methods. Further, the complexity of F layers is higher than that of C layers expect for DIQaM. Especially, in the DB-CNN, RANK, although the number of F layer is much smaller than that of C layer, the complexity of F layer is still higher than the C layer. This is because the F layer optimizes all local features jointly, while the C layer only optimizes local features. Compared with DIQaM and BIECON, since the number FC layers of BIECON methods is much larger than the DIQaM, it is easy to understand that the complexity of BIECON method is higher than the DIQaM. Therefore, F layer has higher effect on DNN complexity than C layer. It is worth noting that when designing the deep network, we need to consider the number of layers and the proportion of C and F layers.

E. DISCUSSION OF DIFFERENT DNN METHODS

As shown in table 10, we compare the implementations and of different DNN methods. The first three DNN models are based on the patch-input methods and the last two DNN methods are based on the image-input methods. Note that SS means image subjective score (SS). DL and DT mean the distortion level (DL) and type (DT), respectively. The comprehensive performance is presented in five different databases (LIVE, TID2013, CSIQ, LIVEMD, LIVEC). In table 10, we find that the prediction performance is not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

TABLE 6: The SROCC and PLCC comparison on the five databases

Types	Algorithms	LIVE		TID2013		CSIQ		LIVEMD		LIVEC		Weighted Average	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
FR	PSNR	0.876	0.872	0.636	0.706	0.806	0.800	0.725	0.815	N/A	N/A	N/A	N/A
	SSIM [3]	0.913	0.945	0.775	0.691	0.834	0.861	0.845	0.882	N/A	N/A	N/A	N/A
	FSIMc [74]	0.963	0.960	0.802	0.877	0.913	0.919	0.863	0.818	N/A	N/A	N/A	N/A
	DeepQA [76]	0.981	0.982	0.961	0.965	0.939	0.947	0.938	0.942	N/A	N/A	N/A	N/A
NR1	BRISQUE [75]	0.939	0.942	0.572	0.651	0.775	0.817	0.897	0.921	0.607	0.645	0.676	0.729
	CORNIA [12]	0.942	0.943	0.549	0.613	0.714	0.781	0.900	0.915	0.618	0.662	0.659	0.708
	GMLOG [13]	0.950	0.954	0.675	0.683	0.803	0.812	0.824	0.863	0.543	0.571	0.713	0.727
	IL-NIQE [14]	0.902	0.908	0.521	0.648	0.821	0.865	0.902	0.914	0.594	0.589	0.651	0.719
	BWS [16]	0.934	0.943	0.597	0.622	0.786	0.820	0.901	0.922	0.482	0.526	0.666	0.693
NR2	MGDNN [35]	0.951	0.949	–	–	–	–	–	–	–	–	–	–
	FRIQUEE [34]	–	–	–	–	–	–	–	–	0.672	0.705	–	–
NR3	GLCP [38]	0.958	0.959	–	–	–	–	–	–	–	–	–	–
	BLNDER [39]	0.966	0.959	0.819	0.838	**0.961**	**0.968**	**0.944**	**0.964**	**0.945**	**0.953**	**0.890**	**0.902**
NR4	DLIQA [43]	0.929	0.934	–	–	–	–	–	–	–	–	–	–
	SESANIN [45]	0.934	0.948	–	–	–	–	0.836	0.838	–	–	–	–
	VPOR [47]	0.967	0.968	–	–	–	–	–	–	–	–	–	–
NR5	CNN [48]	0.956	0.953	0.558	0.653	0.683	0.754	**0.933**	0.927	0.516	0.536	0.604	0.702
	Pre-SM [50]	**0.974**	**0.978**	–	–	–	–	–	–	–	–	–	–
	VIDGIQA [53]	0.969	**0.973**	–	–	–	–	–	–	0.701	–	–	–
	DIQaM [52]	0.960	0.972	0.835	0.855	0.869	0.894	0.906	0.931	0.606	0.601	0.817	0.832
	TSCN [30]	0.969	**0.972**	–	–	–	–	–	–	–	–	–	–
NR6	BIECON [55]	0.961	0.960	0.717	0.762	0.815	0.823	0.909	**0.933**	0.663	**0.705**	0.765	0.797
	DIQA [56]	0.970	0.972	**0.843**	**0.868**	0.844	0.880	0.920	**0.933**	**0.703**	0.704	**0.839**	**0.857**
	BPSQM [57]	0.973	0.963	**0.862**	**0.885**	0.0.874	0.915	–	–	–	–	–	–
NR7	MFIQA [31]	0.964	0.967	–	–	**0.917**	**0.936**	–	–	**0.835**	**0.967**	–	–
	RANk [26]	**0.981**	**0.982**	0.780	0.793	0.892	0.912	0.908	0.929	0.641	0.675	0.800	0.818
	DB-CNN [27]	0.968	0.971	0.816	0.865	**0.946**	**0.959**	**0.927**	**0.934**	**0.851**	**0.869**	**0.868**	**0.897**
	MEON [29]	0.943	0.954	0.808	–	–	–	–	–	–	–	–	–
NR8	HIQA [28]	**0.982**	**0.982**	**0.879**	**0.880**	0.885	0.901	–	–	–	–	–	–

Red: the highest. Blue: the second. Green: the third.

TABLE 7: The RMSE comparison on the five databases

Algorithms	LIVE	TID2013	CSIQ	LIVEMD	LIVEC	Weighted Average
PSNR	12.74	0.567	0.123	8.323	–	–
BRISQUE [75]	9.538	–	–	–	–	–
CORNIA [12]	9.935	–	–	–	–	–
BWS [16]	9.821	0.951	0.192	8.210	22.53	6.479
CNN [48]	7.313	0.921	0.183	7.067	16.38	4.928
RANK [26]	**5.438**	0.818	0.115	5.621	12.25	3.764
BIECON [55]	5.537	**0.452**	0.108	5.365	**6.471**	**2.509**
DIQaM [52]	5.742	0.745	0.120	9.113	15.35	4.595
DIQA [56]	5.793	0.558	**0.114**	**4.960**	10.93	3.391

TABLE 8: The SROCC comparison of the cross dataset test

Train	Test	BRISQUE	GMLOG	BLNDER	CNN	DIQaM	DIQA	RANK	MFIQA
LIVE	CSIQ subset	0.890	0.897	0.700	**0.923**	0.908	0.906	0.797	0.903
	TID2013 subset	0.878	0.907	0.652	**0.920**	0.867	0.918	0.873	–
CSIQ subset	LIVE	0.919	0.903	0.825	–	–	0.923	0.564	**0.933**
	TID2013 subset	0.874	0.879	0.661	–	–	**0.915**	0.777	–
TID2013 subset	LIVE	0.877	0.889	0.751	–	–	**0.905**	0.769	–
	CSIQ subset	0.861	0.794	0.782	–	–	**0.871**	0.735	–

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

TABLE 9: The complexity of different DNN methods

Algorithms	Layers	Input size	Kernel size	Output size	WPs	BPs	ATPs
CNN	C1	$32 \times 32 \times 1$	7×7	$26 \times 26 \times 1$	2450	50	ATPs:7.25×10^5
	F1	$1 \times 1 \times 50(2)$	1×1	$1 \times 1 \times 800$	8000	1600	CPs:2.45×10^4
	F2	$1 \times 1 \times 800$	1×1	$1 \times 1 \times 800$	640000	800	FPs:6.48×10^5
	F3	$1 \times 1 \times 800$	1×1	$1 \times 1 \times 1$	800	1	
DIQaM	C1	$32 \times 32 \times 3$	3×3	$32 \times 32 \times 32$	864	32	ATPs:5.24×10^6
	C2	$32 \times 32 \times 32$	3×3	$32 \times 32 \times 32$	9216	32	CPs:4.71×10^6
	C3	$16 \times 16 \times 32$	3×3	$16 \times 16 \times 64$	18432	64	FPs:5.30×10^5
	C4	$16 \times 16 \times 64$	3×3	$16 \times 16 \times 64$	36864	64	
	C5	$16 \times 16 \times 64$	3×3	$16 \times 16 \times 128$	73728	128	
	C6	$8 \times 8 \times 128$	3×3	$8 \times 8 \times 128$	147456	128	
	C7	$8 \times 8 \times 128$	3×3	$8 \times 8 \times 256$	294912	256	
	C8	$4 \times 4 \times 256$	3×3	$4 \times 4 \times 256$	589824	256	
	C9	$2 \times 2 \times 256$	3×3	$2 \times 2 \times 512$	1179648	512	
	F1	$1 \times 1 \times 512(2)$	1×1	$1 \times 1 \times 512(2)$	524288	1024	
	F2	$1 \times 1 \times 512)$	1×1	$1 \times 1 \times 1$	512	1	
BIECON	C1	$32 \times 32 \times 3$	5×5	$32 \times 32 \times 48$	3648	48	ATPs:7.38×10^6
	C2	$16 \times 16 \times 48$	5×5	$16 \times 16 \times 64$	76864	64	CPs:8.06×10^4
	F1	$8 \times 8 \times 64$	1×1	$1 \times 1 \times 1600$	6553600	1600	FPs:7.294×10^6
	F2	$1 \times 1 \times 1600$	1×1	$1 \times 1 \times 400$	640000	400	
	F3	$1 \times 1 \times 400$	1×1	$1 \times 1 \times 200$	80000	200	
	F4	$1 \times 1 \times 200$	1×1	$1 \times 1 \times 100$	20000	100	
	F5	$1 \times 1 \times 100$	1×1	$1 \times 1 \times 1$	100	1	
RANK	C1	$224 \times 224 \times 3$	3×3	$224 \times 224 \times 64$	1728	64	ATPs:1.34×10^8
	C2	$224 \times 224 \times 64$	3×3	$224 \times 224 \times 64$	36864	64	CPs:1.47×10^7
	C3	$112 \times 112 \times 64$	3×3	$112 \times 112 \times 128$	73728	128	FPs:1.193×10^8
	C4	$112 \times 112 \times 128$	3×3	$112 \times 112 \times 128$	147456	128	
	C5	$56 \times 56 \times 128$	3×3	$56 \times 56 \times 256$	294912	256	
	C6	$56 \times 56 \times 256$	3×3	$56 \times 56 \times 256$	589824	256	
	C7	$56 \times 56 \times 256$	3×3	$56 \times 56 \times 256$	589824	256	
	C8	$28 \times 28 \times 256$	3×3	$28 \times 28 \times 512$	1179648	512	
	C9	$28 \times 28 \times 256$	3×3	$28 \times 28 \times 512$	1179648	512	
	C10	$28 \times 28 \times 256$	3×3	$28 \times 28 \times 512$	1179648	512	
	C11	$14 \times 14 \times 512$	3×3	$14 \times 14 \times 512$	2359296	512	
	C12	$14 \times 14 \times 512$	3×3	$14 \times 14 \times 512$	2359296	512	
	C13	$14 \times 14 \times 512$	3×3	$14 \times 14 \times 512$	2359296	512	
	F1	$7 \times 7 \times 512$	1×1	$1 \times 1 \times 4096$	102760448	4096	
	F2	$1 \times 1 \times 4096$	1×1	$1 \times 1 \times 4096$	16777216	4096	
	F3	$1 \times 1 \times 4096$	1×1	$1 \times 1 \times 1$	4096	1	
DB-CNN	C1	$224 \times 224 \times 3$	3×3	$224 \times 224 \times 48$	1296	48	ATPs:1.85×10^7
	C2	$224 \times 224 \times 48$	3×3	$112 \times 112 \times 48$	20736	48	CPs:5.0×10^5
	C3	$112 \times 112 \times 48$	3×3	$112 \times 112 \times 64$	27648	64	FPs:1.193×10^8
	C4	$112 \times 112 \times 64$	3×3	$56 \times 56 \times 64$	36864	64	
	C5	$56 \times 56 \times 64$	3×3	$56 \times 56 \times 64$	36864	64	
	C6	$56 \times 56 \times 64$	3×3	$28 \times 28 \times 64$	36864	64	
	C7	$28 \times 28 \times 64$	3×3	$28 \times 28 \times 128$	73856	128	
	C8	$28 \times 28 \times 128$	3×3	$28 \times 28 \times 128$	1179648	128	
	C9	$28 \times 28 \times 128$	3×3	$14 \times 14 \times 128$	147456	128	
	F1	$14 \times 14 \times 128$	1×1	$1 \times 1 \times 128$	3211264	128	
	F2	$1 \times 1 \times 128$	1×1	$1 \times 1 \times 256$	9984	256	
	F3	$1 \times 1 \times 256$	1×1	$1 \times 1 \times 1$	9984	39	
	VGG(C1-C13)	–	–	–	14710464	3968	
	F4	$1 \times 1 \times 640$	1×1	$1 \times 1 \times 1$	640	1	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938900, IEEE Access

TABLE 10: The comparison of implementations of different DNN methods

Algorithms	Image size	Train methods	Train label		Network parameters	Performance
			First stage	Second stage		
CNN	32×32	Patch-input	SS	N/A	7.25×10^5	0.702
BIECON	32×32	Patch-input	FSIM	SS	7.38×10^6	0.797
DIQaM	32×32	Patch-input	SS	N/A	5.22×10^6	0.832
RANK	224×224	Image-input	DL	SS	1.34×10^8	0.818
DB-CNN	224×224	Image-input	DL+DT	SS	1.85×10^7	0.897

only related to the complexity of DNN, but also to the strategy of the design algorithm. Although the complexity of DB-CNN is not the highest, the prediction performance is the best in these methods. The reason is that DB-CNN jointly considers three factors. First, they select the image-input method, which can obtain rich distortion information. Second, they consider the distortion types and levels as labels to describe synthetic images in the first stage. Finally, in the second stage, they add a sub-network to predict authentic images.

In addition, since the RANK and DB-CNN methods fix input image size, images need to be cropped or resized as input to meet requirement. It leads to input image is not enough to cover the whole image information and easy to introduce geometric deformation. Therefore, the intermediate label and image size will also be considered to improve prediction performance. Similarly, compared with the patch-input methods, the DIQaM method is superior to others, because the patch saliency is used to solve the inaccurate patch label. Therefore, in order to improve the prediction accuracy, patch size and proxy score will be considered.

In practical application, we need to find a balance between the algorithm complexity and prediction accuracy. For example, in the application of medical images, we pay more attention to the prediction accuracy. On the contrary, in real-time image evaluation system, we will give priority to the algorithm complexity.

V. CHALLENGES OF DNN METHODS

In the previous sections, we present a comprehensive review of the recent literature in DNN models for BIQA. Although DNN-based BIQA methods can achieve outstanding performance due to their strong representation capability, there are several challenges at the same time. Meanwhile, we provide some solutions to these challenges.

1) Creating the large-scale IQA database The number of training samples is critical to the success of DNN models. Currently, the lack of large training data sets is often mentioned as a challenge. Although both the image-input methods and the patch-input methods overcome the problem of insufficient IQA database to some extent, these methods have their own shortcomings to the label accuracy of generation images. Therefore, understanding how to successfully create reliable, very large-scale databases is still an open question.Therefore, the online crowdsourcing system is one possible solution, which aims to gather very rich human data in term of subjective testing. In addition, if a large social media company were to engage their customers to provide image quality scores, it would also ensure the aggregate quality of the collected human data.

2) Exploring unsupervised DNN methods The current DNN models mainly use the supervised end-to-end optimization to evaluate image quality. However, the lack of sufficient ground truth labels is a serious problem for BIQA. Therefore, we expect that training an end-to-end DNN model in a completely unsupervised manner is worth further investigations in the future. This is because obtaining large amounts of unlabeled data is generally much easier than labeled data and human learning is largely unsupervised: we discover the structure of the word by observing it, not by being told the specific labels. Thus,we could try to design two branch networks to the unsupervised method. The one is used to learn the features of reference images and the other is used to learn the distorted images' features. Then, the most important is we need to establish a loss mechanism to quantify the difference between the two branch networks. In addition, the proxy mechanism may be designed to replace the image subjective scores.

3) Explaining the theoretical basis of DNN methods Although DNN thoroughly understands the data distribution and results, for human, there is no theoretical analysis explaining why it works well to the designed DNN architecture and how to further improve the prediction performance. Therefore, it is meaningful to explore the theoretical guarantee of DNN model, in order to guide further researches in this field. The two methods may be selected to explain DNN algorithms. One approach could analyze DNN architecture by using visual method [85]. The visualization of layer-by-layer features helps understand how the DNN learns useful features for IQA task. Another is to explain the functions of DNN according to the algorithms' requirements so that the functions of DNN could deal with the IQA problems.

VI. CONCLUSION

This paper presents a systematic survey of various DNN-based methods for BIQA. We discussed and analyzed the state-of-the-art DNN methods according to different strategies of DNN models. This classification strategy explicitly shows the characteristics, advantages and disadvantages of different DNN methods for BIQA. Especially, some novel

DNN methods, which are not present in previous study, are also discussed. Then, we compare the performance and complexity of various DNN models, yet the state of research in this field is far from mature. Meanwhile, we summarize the intrinsic relationship among different DNN methods and obtain some interesting findings, which can help us design DNN for BIQA. Furthermore, we provide several challenging issues of using DNN methods for BIQA, which should be noticed. We hope this survey of DNN methods can serve as a useful reference towards a better understanding of this research field.

REFERENCES

[1] F. Li, S. Fu, Z. Li and X. Qian, "A cost-constrained video quality satisfaction study on mobile device," IEEE Trans. Multimedia, vol. 20, no. 5, pp. 1154–1168, May. 2018.

[2] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma and W. Gao, "Fine-grained quality assessment for compressed image," IEEE Trans. Image Process., vol. 28, no. 3, pp. 1163–1175, Apr. 2019.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[4] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," IEEE Trans. Image Process., vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[5] H. Jia, L. Zhang and T. Wang, "Contrast and visual saliency similarity-induced index for assessing image quality," IEEE Access, pp. 65885–65893, 2018.

[6] S. A. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes," IEEE Trans. Image Process., vol. 25, no. 11, pp. 5293–5303, Nov. 2016.

[7] S. Wang, K. Gu, X. Zhang, W. Lin, L. Zhang, S. Ma and W. Gao, "Subjective and objective quality assessment of compressed screen content images," IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 6, no. 4, pp. 532–543, Dec. 2016.

[8] Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao and W. Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," IEEE Trans. Multimedia, vol. 20, no. 2, pp. 379–391, 2018.

[9] Y. Ding, R. Deng, X. Xie, X. Xu, Y. Zhao, X. Chen and A. S. Krylov, "No-reference stereoscopic image quality assessment using convolutional neural network for adaptive feature extraction," IEEE Access, pp. 37595–37603, 2018.

[10] L. Li, Y. Yan, Q. Lu, J. Wu, K. Gu and S. Wang, "No-reference quality assessment of deblurred images based on natural scene statistics," IEEE Access, pp. 2163–2171, 2017.

[11] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," in Proc. CoRR, 2017.

[12] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in Proc. IEEE Conf. CVPR, pp. 1098–1105, Jun. 2012.

[13] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," IEEE Trans. Image Process., vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[14] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," IEEE Trans. Image Process., vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[15] Q. Wu, H. Li, F. Meng, K. Ngan, B, Luo, C. Huang and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," IEEE Trans. Circuits and Systems for Video Technology, pp. 425–440, 2016.

[16] X. Yang, F. Li, W. Zhang and L. He, "Blind image quality assessment of natural scenes based on entropy differences in the DCT domain," Entropy, vol. 20, no. 12, pp. 885–906, 2018.

[17] A. Krizhevsky, I. Sutskever, and H. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. NIPS, pp. 1097–1105, 2012.

[18] C. Szegedy, W. Liu and Y. Jia, "Going deep with convolutions," in Proc. IEEE Conf. CVPR, pp. 1–9, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. CVPR, pp. 770–778, 2016.

[20] K.Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computer Science, pp. 1–14, 2014.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma and Z.Huang,"ImageNet large scale visual recognition challenge," in Proc. IJCV., 2015.

[22] R. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: a survey," Information Sciences, vol. 301, pp. 141–160, 2015.

[23] A. Ebrahimi Moghadam, P. Mohammadi, and S. Shirani, "Subjective and objective quality assessment of image: A survey," Majlesi J. Elect. Eng., vol. 9, pp. 1–50, Mar. 2015.

[24] A. G. George and K. Prabavathy, "A survey on different approaches used in image quality assessment," Int. J. Computer Sci. and Network Security, vol. 14, no. 2, pp. 197–203, 2014.

[25] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," IEEE Signal Processing Magazine, pp. 130–141, 2017.

[26] X. Liu, J. Weijer, and A. Bagdanov, "RankIQA: Learning from ranking for no-reference image quality assessment," in Proc. IEEE Conf. ICCV, pp. 1040–1049, 2017.

[27] W. Zhang, K. Ma, J. Yan, D. Deng and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," IEEE Trans. Circuits and Systems for Video Technology,2018.

[28] K. Lin and G. Wang, "Hallucinated-IQA:no-reference image quality assessment via adversarial learning," in Proc. IEEE Conf. CVPR, pp. 732–741, Aug. 2018.

[29] K. Ma, W. Liu, Z. Duanmu, Z. Wang and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," IEEE Trans. Image Process., vol. 27, no. 3, pp. 1202–1213, 2018.

[30] Q. Yan, D. Gong, and Y. Zhang, "Two-stream covolutional networks for blind image quality assessment," IEEE Trans. Image Process., vol. 28, no. 5, pp. 1202–1213, 2019.

[31] J. Kim, A. Ngugen, S. Ahn, C. Luo, and S. Lee, "Multiple level feature-based blind image quality assessment model," in Proc. IEEE Conf. ICIP, pp. 291–295, 2018.

[32] X. Yang, F. Li, and H. Liu, "A comparative study of DNN-based models for blind image quality prediction," in Proc. IEEE Conf. ICIP, 2019.

[33] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in Proc. IEEE Conf. CVPR., pp. 2877–2884, 2014.

[34] D. Ghadiyaram and A. C. Bovik, "Blind image quality assessment on real distorted images using deep belief nets," in Proc. IEEE Global Conf. Signal Inf. Process., pp. 946–950, 2014.

[35] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, and S. Liu, "Difference of Gaussian statistical features based blind image quality assessment: A deep learning approach," in Proc. IEEE Conf. ICIP., pp. 2344–2348, 2015.

[36] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation., vol. 18, no. 7, pp. 1527–1557, 2006.

[37] D. Li, N. T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in Proc. ACM MM, pp. 378–386, Oct, 2017.

[38] C. Sun, H. Li and W. Li, "No-reference image quality assessment based on global and local content perception," in Proc. IEEE Conf. VCIP, pp. 27–30, 2016.

[39] F. Gao, J. Yu, S. Zhu, Q. Huang and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," Patter Recognition, vol. 81, pp. 432–442, 2018.

[40] K. Gu, G. Zhai, X. Yang and W. Zhang, "An efficient color image quality metric with local-tuned-global model," in Proc. ICIP, pp. 506–510, 2014.

[41] N. Kruger, P. Janssen, S. Kalkan, M, Lappe and A. Leonardis, "Deep hierarchies in the primate visual cortex: what can we learn for computer vision?," IEEE Trans. Pattern Anal. Mach., pp. 1847–1871, 2013.

[42] E. Hutchins and G. Lintern, Cognition in the Wild. Cambridge., MA, USA: MIT Press, 1995.

[43] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," IEEE Trans. Neural Netw. Learn. Syst.,vol. 26, no. 6, pp. 1275–1286, 2015.

[44] W. Hou and X. Gao, "Saliency-guided deep framework for image quality assessment," IEEE Multimedia Mag.,vol. 22, no. 2, pp. 46–55, 2015.

[45] Y. Li, L.-M. Po, X. Xu, L. Feng, F. Yuan, C.-H. Cheung, and K.-W. Cheung, "No-reference image quality assessment with Shearlet transform and deep neural networks," Neurocomputing,vol. 154, pp. 94–109, 2015.

[46] S. Bianco, L. Celona. P, Napoletano and R. Schettini, "On the use of deep learning for blind image quality assessment," Signal Image Video Processing, pp. 355–362, 2017.

[47] J. Gu, G. Meng, J. A. Redi and C. Pan, "Blind image quality assessment via vector regression and object oriented pooling," IEEE Trans. Multimedia, vol. 20, no. 5, pp. 1140–1153, 2018.

[48] L. Kang, P. Ye, Y. Li and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in Proc. IEEE Conf. CVPR, pp. 1733–1740, 2014.

[49] C. Pan, Y. Xu, Y. Yan, K. Gu and X. Yang, "Exploiting neural models for no-reference image quality assessment," in Proc. VCIP, pp. 27–30, 2016.

[50] Z. Cheng, M. Takeuchi and J. Katto, "A pre-saliency map based blind image quality assessment via convolutional neural networks," in proc. IEEE International Symposium on Multimedia, pp. 77–82, 2017.

[51] S. Jia and Y. Zhang, "Saliency-based deep convolutional neural network for no-reference image quality assessment," Multimedia Tools and Application, 2017.

[52] S. Bosse, D. Maniry, K. Muller, T. Wiegand and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," IEEE Trans. Image Process., vol. 27, no. 1, pp. 206–219, 2018.

[53] J. Guan, S. Yi, X. Zeng, W. Cham and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," IEEE Trans. Multimedia, vol. 19, no. 11, pp. 2505–2521, 2017.

[54] B. Bare, K. Li and B. Yan, "An accurate deep convolutional neural networks model for no-reference image quality assessment," in Proc. ICME, pp. 1356–1361, 2017.

[55] J. Kim and S. Lee, "Fully deep blind image quality predictor," IEEE J. Sel. Topics Signal Process., vol. 11, no. 1, pp. 206–220, Feb. 2017.

[56] J. Kim, A. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," IEEE Trans. Neural Netw. Learn. Syst., pp. 1–14, 2018.

[57] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in Proc. IEEE Int. Conf. CVPR, 2018.

[58] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Visual saliency detection with free energy theory," IEEE Singal Processing Letter, vol. 22, no. 10, pp. 1552–1555, 2015.

[59] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: efficient and reliable full-reference image quality evaluator," IEEE Access, pp. 5579–5590, 2016.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015.

[61] Y. Li, L. M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in Proc. IEEE Int. Conf. Digital Signal Processing, pp. 685–689, 2016.

[62] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," In Advances in neural information processing systems, pp. 487–495, 2014.

[63] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowledge and Data Engineering, pp. 1345–1359, 2010.

[64] M. Lin, Q. Chen, and S. Yan, "Network in network," in Proc. ICLR, 2014.

[65] H. Sheikh, M. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process., vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[66] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K.Chehdis, M. Carli, F. Battisti, C. C. Kuo, "Image database TID2013: Peculiarities, results and perspectives," Signal Process. Image Commun., pp. 57–77, Jan. 2015.

[67] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full reference image quality assessment and the role of strategy," J. Electron. Imag., vol. 19, no. 1, pp. 19–21, 2010.

[68] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in Proc. Asilomar Conf. Signals, Syst. Comput., pp. 1693–1697, 2012.

[69] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo Exploration Database: New challenges for image quality assessment models," IEEE Trans. Image Process., vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[70] S. Chopra, R. Hadsell, and Y. Cun, "Learning a similarity metric discriminatively with application to face verification," in Proc. IEEE Conf. CVPR, pp. 349–356, 2005.

[71] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[72] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," IEEE Trans. Image Process., vol. 25, no. 1, pp. 372–387, Jan. 2016.

[73] A. M. Rohaly et al, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, ITU-T Standards Contrib. COM, pp. 9–80, 2000.

[74] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "MassFSIM: A feature similarity index for image quality assessment," IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[75] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process., vol. 21, no. 12, pp. 4695–4708, 2012.

[76] J. Kim and S. Lee, "Deep learning of human visual sensitivity in FR-IQA framework," in Proc. CVRP, pp. 1676–1684, 2017.

[77] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," J. Vision, vol.17, no.1, pp. 32–59, 2017.

[78] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," IEEE Trans. Cybern., vol.46, no.1, pp. 39–50, 2016.

[79] H. Liu, N. Klomp, I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," IEEE Trans. Circuits Syst.Video Technol., vol.20, no.4, pp. 529–539, 2010.

[80] W. Ji, J. Wu, M. Zhang, G. Shi and X. Xie, "Blind image quality assessment with joint entropy degradation," IEEE Access, pp. 30925–30936, 2019.

[81] Q. Wu, H. Li, K. Ngan and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," IEEE Trans. Circuits Syst.Video Technol., vol.28, no.9, pp. 2078–2089, 2018.

[82] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang and X. Gao, "No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness," IEEE Trans. Image Process., 2019.

[83] J. Wu, J. Zeng, W. Dong, G. Shi, and W. Lin, "Blind image quality assessment with hierarchy: Degradation from local structure to deep semantics," J. Visual Communication and Image Representation, pp. 353–362, 2019.

[84] F. Meng, L. Guo, Q. Wu, and H. Li, "A New Deep Segmentation Quality Assessment Network for Refining Bounding Box Based Segmentation," IEEE Access, vol.7, pp. 59514–59523, 2019.

[85] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. ECCV, 2014.

[86] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in Proc. CVPR, 2011.

[87] N. Joshi, C. L. Zitnick, R. Szeliski, and D. Kriegman, "Image deblurring and denoising using color priors," in Proc. CVPR, 2009.

[88] N. Joshi, R. Szeliski, and D. Kriegman, "Psf estimation using sharp edge prediction," in Proc. CVPR, 2009.

XIAOHAN YANG obtained her M.S. degree in Information Engineering and Automation from Kunming University of science and technology, Kunming, China, in 2016. She is currently pursuing the Ph.D. degree at School of Electronic and Information Engineering, Xi'an Jiaotong University. Her research interests mainly focus on image quality assessment.

FAN LI obtained his B.S. and Ph.D. degrees in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2003 and 2010, respectively. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of California, San Diego. He is currently a Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University. He has published more than 30 technical papers. His research interests include multimedia communication and video quality assessment. He served as the Local Chair for ICST Wicon 2011, and was a member of the Organizing Committee for IET VIE 2008.

HANTAO LIU received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Assistant Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is serving for the IEEE MMTC, as the Chair of the Interest Group on Quality of Experience for Multimedia Communications, and he is an Associate Editor of the IEEE Transactions on Human-Machine Systems and the IEEE Transactions on Multimedia.